STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Room Paper No. 2
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

IMPACT OF NEW TECHNOLOGICAL DEVELOPMENTS IN SOFTWARE, COMMUNICATIONS AND COMPUTING ON SDC:

List of key issues for discussion

Josep Domingo-Ferrer
Department of Computer Engineering & Mathematics
Universitat Rovira i Virgili
Autovia de Salou, s/n
E-43006 Tarragona, Catalonia
e-mail jdomingo@etse.urv.es

Abstract

An overview of papers submitted on Topic 2 (Impact of new technogical developments in software, communications and computing on SDC) is first given. Then a list of key issues for substantive discussion are identified.
Keywords: Statistical data protection, Statistical data confidentiality, Software for SDC, Data security, Official statistics.

I. INTRODUCTION

1. There have been eleven contributed papers from seven different countries submitted on Topic II (Impact of new technological developments in software, communications and computing on SDC). This number of (non-invited) contributions clearly reflects the high level of current research and development activity in this topic. From the point of view of their contents, those eleven papers split into five thematic blocks:

- *Systems for access to microdata.* The paper by Tambay and White (Paper no. 13) describes three approaches developed by Statistics Canada to provide researchers with access to data from complex surveys: the first approach is based on public-use microdata files, the second on remote access and the third on research data centers. In Paper no. 21, Hawala outlines the operation of American FactFinder (AFF), the U.S. Census Bureau's new online dissemination system; the system provides tabular data build on microdata from several censuses and surveys. An outstanding AFF feature is that users can define their own tables (what is called "Tier 3 access"), which is a real SDC challenge.
- *Tabular protection.* In Paper no. 16, Salazar presents a new methodology called Partial Cell Suppression for the Cell Suppression Problem (CSP). Partial CSP replaces suppressed cells with intervals rather than "deleting" such cells. The idea originates from the fact that "deleting" a cell

actually amounts to replacing it with the feasibility interval that can be computed from marginals using an LP-solver. Indeed, partial CSP can lead to replacing more cells by intervals than complete CSP, but the overall information loss can be lower; in addition partial CSP appears to be easier than complete CSP from a computational standpoint. A combination of complete CSP and partial CSP is proposed by the author of Paper no. 16. The paper by Castro and Heredia (Paper no. 24) discusses the use of modelling languages for quick development of algorithm prototypes for CSP; prototyping is illustrated with a particular network flow method, for which some preliminary computational results are presented.

- *Reidentification*. In Paper no. 16, Torra explores a novel and very realistic reidentification scenario, namely record linkage between two datasets (the original dataset and the SDC-protected dataset) when the variables known by the intruder in both datasets are not exactly the same but similar. Reidentification procedures based on clustering are introduced which give encouraging results. Reidentification for the specific case of sampling methods for microdata protection is discussed by Elliot in Papers no. 19 and 20. Paper no. 19 explains the special uniques identification method. Paper no. 20 describes the DIS (Data Intrusion Simulation) method for calculating the general disclosure risk for a given target file, without using population statistics nor matching experiments for record linkage.
- *Microdata protection methods*. Paper no. 14 by Tammilehto-Luode presents the data protection guidelines proposed by an internal workgroup at Statistics Finland; a new protection method aimed at protecting geographically detailed data is also outlined. Paper no 22 by Franconi and Stander introduces a new microdata protection paradigm (model-based paradigm) which should outperform microaggregation for releasing geographically referenced data. Experiences on model-based protection are reported in Paper no. 15 by Franconi, Capobianchi, Polettini and Seri.
- *SDC software.* Paper no. 23 by Hundepool delineates the objectives and the approach of the EU-funded CASC project, which should yield (among other deliverables) an improved version of the Argus SDC software.

2. Section II of this document contains a list of key questions for substantive discussion on the microdata access subtopic. Section III contains a list of key questions on the tabular protection subtopic. Section IV deals with questions on reidentification. Possible discussion issues on the microdata protection subtopic are listed in Section V. Section VI is on SDC software. Section VII is a conclusion.

II. LIST OF KEY QUESTIONS ON SYSTEMS FOR ACCESS TO MICRODATA

3. In [Paper13], the approaches of Statistics Canada to providing access to survey data are explained. Public Use Microdata Files (PUMF), remote access and research data centers are explored in turn. Some questions related to this paper follow:

- The approach to assessing disclosure risk for PUMFs is a very pragmatic one, namely record matching. The authors mention using nearest neighbour-matching, where the distance is based on the largest univariate relative difference. Why these choices? Would other distances (such as the multidimensional Euclidean or the average univariate relative difference) not appear as more natural and/or robust? Regarding the matching algorithm, have sophisticated proposals such as [Jaro89] been considered? Other statistical offices (e.g. USCB) have been using [Jaro89] in their matching experiments. In general, more details on the security assessment of the PUMF being released would be welcome.
- Remote access as discussed in the paper is not an automated procedure. There are currently initiatives toward automating remote access (see [Paper21]). Is this automation perceived as necessary and feasible? If so, how to automatically check if a query is answerable without infringing statistical confidentiality?
- Regarding research data centers, is encryption being used for on-line communication between them and Statistics Canada central premises? If not, how are those centers fed with current microdata?

4. U.S. Bureau of the Census's American FactFinder (AFF) tool is described in [Paper21]. AFF is a bold initiative toward providing on-line customized access to microdata resources. In its current state, two tiers of data access are defined: under Tier 2 access the Census defines the tables and users can choose from the corresponding list of tables; under Tier 3 access users can define their own tables. It is stated in the paper that summarized data from Tier 3 will be provided only if they pass disclosure limitation rules. A query filter is supposed to detect queries that will not pass such disclosure rules before they are submitted; the query filter is complemented by a results filter which performs a final check on the resulting table before returning it to the user.

- One of the rules used by the query filter relates to the estimated time or size of the output exceeding a predefined Census threshold. It would be interesting to know some details on how such estimate can be obtained.
- A second, more general, question is how does the system handle repeated queries, i.e. how does it guard against the tracker attack [Schlö80]. In fact, both the query and the results filter should take into account previous queries by the same user before allowing a certain query to be answered. Otherwise, an intruder can accumulate knowledge through successive queries and eventually succeed in reidentifying an individual. This is the well-known weak point of on-line statistical databases and countermeasures are not obvious.

III. LIST OF KEY QUESTIONS ON TABULAR PROTECTION

5. In [Paper16] a new methodology called partial cell suppression is presented for the Cell Suppression Problem (CSP). Partial cell suppression replaces suppressed cells with intervals rather than "deleting" such cells. The idea originates from the fact that "deleting" a cell actually amounts to replacing it with the feasibility interval that can be computed from marginals using an LP-solver. Indeed, partial cell suppression can lead to replacing more cells by intervals than complete cell suppression, but the overall information loss can be lower; in addition partial CSP appears to be easier than complete CSP from a computational standpoint. In view of this:

- Why is it necessary to combine partial cell suppression with complete cell suppression? If partial cell suppression loses less information and is computationally easier, what is the advantage of combining it with complete cell suppression?
- There is an argument that partial cell suppression replaces more cells by intervals than complete cell suppression. If the intervals resulting from partial cell suppression are narrower, could not a few more intervals be acceptable?
- Is there some connection between partial cell suppression and the camouflage technique proposed for microdata in [Gopa99]? In the latter case, microdata were systematically replaced by interval answers resulting from an optimization problem.

6. Modelling languages are presented in [Paper24] for quick development of algorithm prototypes for CSP; prototyping is illustrated with a particular network flow method, for which some preliminary computational results are presented.

- Aside from the illustration purpose, can network flow algorithms provide heuristics that can favourably compare to iterative procedures like USBCSUP and GHQUAR [Gieβ98] in the case of big tables?
- Could perhaps this class of algorithms bridge the gap between exact algorithms [Paper16] and iterative procedures, in the sense of dealing with medium-sized tables?

IV. LIST OF KEY QUESTIONS ON REIDENTIFICATION

7. In [Paper16] record linkage between two datasets (the original dataset and the SDC-protected dataset) is considered when the variables known by the intruder in both datasets are not exactly the same but similar. Record matching procedures based on clustering are introduced which give interesting results. The concept of "structural information" is a very powerful one and can be a serious threat to microdata SDC as it is understood today. In fact the two datasets being matched do not need

to share any variables; it suffices that there are variables in both datasets which have a similar structure.

- In the example, two sets of 175 records are needed to reidentify 34 objects (variables). In fact, the author reidentifies variables because there are not enough instances to reidentify individuals. In general, if two datasets correspond to a group of *n* individuals, how many records would be needed for this procedure to be successful? Is not the need for more records than individuals a hindrance to using this approach to attack SDC methods? If so, what possible improvements can be envisioned?
- Are there any plans to compare the resistance of the various SDC masking methods against this matching approach?

8. Reidentification for the specific case of sampling methods for microdata protection is discussed by Elliot in [Paper19,Paper20]. Paper no. 19 explains the special uniques identification method. Paper no. 20 describes the DIS (Data Intrusion Simulation) method for calculating the general disclosure risk for a given target file, without using population statistics nor matching experiments for record linkage.

- Both papers focus on reidentification for sampling methods. What about perturbative masking? Could some of the strategies discusses still be adapted?
- More generally, what are the reasons (if any) for preferring sampling methods to perturbative methods?

V. LIST OF KEY QUESTIONS ON MICRODATA PROTECTION METHODS

9. The data protection guidelines proposed by an internal workgroup at Statistics Finland are listed in [Paper14]; the paper also describes a new protection method aimed at protecting geographically detailed data. [Paper22] introduces a new microdata protection paradigm (model-based paradigm) which should outperform microaggregation for releasing geographically referenced data. Experiences on model-based protection are reported in [Paper15].

- [Paper14], [Paper22] and [Paper15] report on two SDC methods for protection of geographically referenced data. Both proposals are based on imputation. Are there other similarities that can be exploited? Possibly there is room for joint work.
- [Paper22] mentions that neither independent noise, nor microaggregation nor data swapping are suitable for the protection of economic data. Therefore the authors use a new approach, model-based protection. Have other existing methods been tried? For example, rank swapping [Moor96] seems to perform well in the comparison [Paper5].
- What are the information loss and the disclosure risk metrics used to state that model-based protection performs better? Are information loss metrics dependent on the particular data use? If so, which data uses have been considered?

VI. LIST OF KEY QUESTIONS ON SDC SOFTWARE

10. The objectives and the approach of the main EU-funded initiative on SDC, the CASC project, are explained in [Paper23]. The most visible outcome of CASC is an improved version of the Argus software, including protection for large and complex tables and a comprehensive array of microdata masking algorithms.

- Would it be interesting to define common data sets and common software evaluation criteria? Such need was identified at SDP'98 [Domi99] and again at the last Joint UNECE/Eurostat Work Session. The availability of common data sets, common information loss metrics and common disclosure metrics would make it easier to find the most suitable method for a given data type or use. This applies both to tabular and to microdata protection.

VII. CONCLUSIONS

11. Research in SDC is more necessary than ever. Recent challenges which require more powerful SDC protection are the following:

(i) The need to provide customized access to statistical resources is appearing as an unavoidable challenge to statistical institutes. On-line systems offering automated or semi-automated disclosure control are starting to appear. The American FactFinder system is a pioneering initiative in this field. On the other hand, on-line query systems are more vulnerable to iterative attacks (e.g. trackers).

(ii) Record matching and data mining are rapidly evolving fields. As an example, Paper no. 17 [Paper17] proposes an algorithm for record matching between two datasets which do not even need to share the same variables. SDC methods must take those advances into account and provide adequate protection. Ultimately, disclosure risk can be viewed as matching risk.

(iii) Disseminating geographically referenced microdata is another common concern of statistical offices. Current microdata protection methods may not be specially designed for that mission. This should be examined and new methods should be developed if necessary.

12. Joint research efforts like the CASC project are indeed necessary to reach the critical mass to successfully tackle the above challenges.

References

[Paper24] J. Castro and F. J. Heredia, Using modelling languages for the complementary suppression problem through network flow models, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 24, 2001.

[Paper5] J. Domingo-Ferrer and J. M. Mateo-Sanz, An empirical comparison of SDC methods for continuous microdata in terms of information loss and disclosure risk, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 5, 2001.

[Domi98] J. Domingo-Ferrer and J. M. Mateo-Sanz, Current directions in statistical data protection-Preface to SDP'98, *Research in Official Statistics*, no. 2, pp. 105-112, 1998.

[Paper19] M. Elliot, The identification of special uniques, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Paper no. 19, 2001.

[Paper20] M. Elliot, Data intrusion simulation: advances and a vision for the future of disclosure control, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Paper no. 20, 2001.

[Paper15] L. Franconi, A. Capobianchi, S. Polettini and G. Seri, Experiences on model-based disclosure limitation, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 15, 2001.

[Paper22] L. Franconi and J. Stander, Microaggregation and model-based methods for disclosure limitation and their application to business microdata, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* Paper no. 22, 2001.

[Gieβ98] S. Gieβing, Looking for efficient automated secondary cell suppression systems: a software comparison, *Statistical Data Protection'98,* Luxembourg: Office for Official Publications of the European Communities, 1999.

[Gopa98] R. Gopal, P. Goes and R. Garfinkel, Confidentiality via camouflage: the CVC approach to database query management, *Statistical Data Protection'98,* Luxembourg: Office for Official Publications of the European Communities,1999.

[Paper21] S. Hawala, American FactFinder: US Bureau of the Census works towards meeting the needs of users while protecting confidentiality, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 21, 2001.

[Paper23] A. Hundepool, The CASC project, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 23, 2001.

[Jaro89] M. A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, *Journal of the Americal Statistical Association*, vol. 84:414-420, 1989.

[Paper18] S. Karajovanovic and V. Dzukeska, Statistical data protection in the State Statistical Office: technical aspects, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 18, 2001.

[Moor96] R. Moore, Controlled data swapping techniques for masking public-use microdata sets, U.S. Bureau of the Census, working paper, 1996.

[Paper16] J.-J. Salazar, Improving cell suppression in statistical disclosure control, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 16, 2001.

[Schlo80] J. Schlörer, Disclosure from statistical databases: quantitative aspects of trackers, *ACM Transactions on Database Systems*, vol. 4(1), pp. 467-492, 1980.

[Paper13] J. L. Tambay and P. White, Providing greater accessibility to survey data for analysis, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 13, 2001.

[Paper14] M. Tammilehto-Luode, Disclosure control for demographic statistics – Redefined guidelines and development of methods at Statistics Finland, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 14, 2001.

[Paper17] V. Torra, On the re-identification of individuals described by means of non-common variables: a first approach, *2nd Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Paper no. 17, 2001.